An approach to compare Affymetrix and cDNA gene expressions

Chintanu Kumar Sarmah* and Sandhya Samarasinghe
Centre for Advanced Computational Solutions (C-fACS), Lincoln University, Christchurch, New Zealand

The development of microarray technology for high-throughput measurement of gene expressions is proving a powerful means for studying the transcriptome on a genomic scale and across diverse biological systems and experimental designs. The technology has grown rapidly in academia, medicine and the pharmaceutical, biotechnology, agrochemical and food industries. The technology confers the freedom to conduct experiments in its multiple platforms and attends to the extensive molecular surveillance of cells and tissues. Despite a large number of studies available in the literature comparing results between the existing platforms remains a challenging task in microarray technology. In this direction, we have developed an approach that allows integration of microarray experiments between two of its platforms, i.e. Affymetrix® and cDNA microarrays. In this communication, we elaborate on our approach and present our validation. In addition, we have also examined our approach to evaluate where it stands amid a few existing methods.

Keywords: Affymetrix, cDNA, gene expression, integration, microarray.

The new millennium is currently witnessing a high-paced information revolution that was initiated in the latter part of the 20th century. In 1995, the technique of microarrays was introduced1,2 and along with it launched the era of simultaneous interrogation of thousands of genes. This was a ground-breaking innovation catalysing a sudden transformation of approach to biological research from the conventional way of working on gene-by-gene basis. Over the relatively short span of time since its inception, microarray technology has improved dramatically and positioned itself as a widely used and reliable high-throughput tool for studying global gene expression while addressing increasingly complex biological questions.

From the late nineties, investigators have started conducting microarray experiments using either of the two distinct techniques – cDNA microarrays and oligonucleotide microarrays. With the development of this field, different laboratories too have begun to routinely make customized arrays. All these have overall contributed immensely to the technological and conceptual advancement. Nevertheless, there are issues in this domain that remain yet unresolved, which include reliability and reproducibility3,4, experimental design1, statistical issues5,6,7,8,9,10 and others5,11. One such critically unresolved niche of microarray technology lies in the integration of data from different microarray experiments.

There is incomparability between cDNA and oligonucleotide-based Affymetrix® platforms; and in this direction, we present an approach that accounts for the differences in both the platforms while delivering a way of comparison of the data. Using childhood leukaemia data (available online at http://tinyurl.com/26k46r2) supplied by Tumour Bank, The Children's Hospital at Westmead, Australia, we demonstrate the utility of the approach, which addresses the incomparability of the two microarray platforms, and subsequently, brings the two to a comparable level. This communication elaborates on the proposed approach and presents our validation in this regard. Further, we have also probed into the approach to evaluate where it stands amid a few existing methods from the literature.

Affymetrix GeneChip® and GenePix® cDNA data belonged to childhood leukaemia patients are used in this work, and seven of these patients were tested both in Affymetrix (HGU-133A chip) and cDNA platforms. Additionally, there are 10 healthy children tested on Affymetrix (HGU-133A) platform. Once the data are obtained, an elaborate quality assessment of these data is carried out using open-source statistical software, R (ref. 11) and bioconductor12. This is followed by normalization, which is employed to the quality-assessed data to minimize systematic variations in the measured gene expression levels.

Quantile normalization method using Robust Multichip Average or RMA-algorithm13 is found relatively effective for normalizing the group of 17 Affymetrix chips. Prior to normalizing the cDNA arrays, an adaptive background correction, Normexp + offset (where, offset = 50) is used, as recommended by Ritchie et al.14. Then, the location-based loess normalization method, Printtiploess15,16, is applied as it performs best in normalizing the GenePix (Axon Instruments, Redwood City, CA, USA)-generated cDNA arrays.

In the cDNA platform, the problem of noise is higher than in the platforms like Affymetrix because the former has more scope for noise to be introduced from the stage of array-construction up to scanning of the images. It is also reported by Lee et al.17 that in cDNA, the probability that a single spot will display as a signal even if the mRNA is not present is as large as 10%, whereas non-displaying of a signal while a spot does contain complementary DNA remains at a non-negligible probability of about 5%. Moreover, in comparison to the oligonucleotide libraries, there are concerns involving the probe contents of cDNA libraries about annotation, clone identity and probe performance18. The Affymetrix platform too has issues, such as non-specific hybridization and less

*For correspondence. (e-mail: sarmahc3@lincoln.ac.nz)
than optimal choice of the oligonucleotide sequences representative of a gene. Nevertheless, the concerns with cDNA arrays often come up more predominantly, contributing to the fact that they have issues with reliability and that the differentially expressed (DE) genes do not necessarily match in identical microarray studies. In this context, the present study relies more on the normalized Affymetrix arrays to select the list of DE genes. The same set of genes from the cDNA platform is then extracted to be considered as the genes of interest for this platform.

Like most other array manufacturers, Affymetrix includes a number of control probes on their arrays. For finding the DE genes, a set of 68 of such control probes is removed. Then, the low quality data are filtered out based on the fact that the probe sets scoring absent or marginal can be considered suspect, whereas present scores are good indicators of signal reliably above background noise. To get rid of data with low content of information, the relative standard deviation, also known as the coefficient of variability (CV), is used. It filters out the least variable genes, defined by the 90th percentile of the distribution of CV values. Figure 1 is constructed by plotting the negative logarithm of the $P$-values as a function of the base 2 log-transformed fold changes. Here, the statistically significant genes are highlighted with sharp circles and 2-fold limits are symbolized by vertical lines. The statistical significance cut-off (0.05) is overlaid as a horizontal line.

The filtering out of the uninteresting genes reduced the total number of highly variable genes to 2,222, from the original number of 22,283.

To the shortened list of genes belonging to the two experimental conditions (healthy and leukaemic), an empirical Bayes method is applied. This is an adaptive strategy towards increasing statistical power and simultaneously reducing the risk of false positives. The $P$-values, subsequently obtained, are adjusted to account for the multiple testing (or, multiple comparisons) problem. It is done by using the method of Benjamini and Hochberg that controls the false discovery rate (FDR), the expected proportion of the significant results that are in fact type-I errors (false discoveries) amongst the rejected hypotheses in multiple comparisons. The FDR is a relaxed condition; and the Benjamini and Hochberg’s method is a better compromise between sensitivity and specificity as it controls the proportion of false significant results instead of controlling the chance of making even a single type-I error. For our data, FDR control is set to a conservative value of 0.05.

The overall procedure on Affymetrix expression levels picks a total of 822 genes as DE. UniGene database is used to annotate these genes. The same set of genes is then considered as differentially expressed in the cDNA platform as well; and accordingly, their expression levels are retrieved. It is assumed that as the arrays in both the platforms belong to the same leukaemia patients; the same set of genes would ideally be expressed differentially in either platform.

A fact for Affymetrix and cDNA data is that they invariably do not hold any relationship between them at all. This holds true here too with regard to the common patients tested on both the platforms. The data obtained
for these patients from the two platforms bear absolutely no relationship. Then, while using the retrieved DE genes, the correlation (Pearson product–moment correlation coefficient, \( r \)) between the data from both the platforms is found to be 0.13, indicating a marginal improvement.

The set of 822 DE genes detected in leukaemic children is also identified in each of the 10 healthy arrays. The RMA normalization produced log2 expressions for all Affymetrix arrays, i.e. for both healthy and leukaemic patients. The log2 expression values of 822 DE genes of the healthy Affymetrix arrays are converted to their respective anti-logs, and the expression value of each gene is averaged across the 10 healthy arrays. It gives rise to a single, averaged and log-free expression for each of the 822 genes. Simultaneously, the expression antilogs of DE genes belonging to the seven-leukaemic patients are calculated. Then, the Affymetrix-ratio or Affy\(_{\text{ratio}}\) for a gene of a patient is estimated by dividing the calculated expression by the average antilog of the corresponding gene from the healthy Affymetrix arrays and subsequent log2 conversion of the obtained value. If expression level of a gene, \( x \) from one of the leukaemic Affymetrix chips is \( D \) and the average expression of each gene from the set of 10 healthy Affymetrix chips is \( H \), then the Affy\(_{\text{ratio}}\) is denoted by eq. (1). It assures that similar to cDNA, where the expression level of a gene remains in the form of a tumour-to-healthy ratio, this transformation converts the Affymetrix expression data into tumour-to-healthy ratios.

\[
\text{Affy}_{\text{ratio}} = \log_2 \frac{\text{Anti log}(D_x)}{\frac{\sum_{x=1}^{10} \text{Anti log}(H_x)}{10}}
\]

With this changeover, both cDNA and Affy\(_{\text{ratio}}\) data can be, in theory, considered to attain a mutually comparable level. Subsequently, the correlation between Affy\(_{\text{ratio}}\) and cDNA is tested, and DE is found to have now increased to 0.6, which is approximately a 6-fold improvement from the previously obtained result. Further, the distribution of the original Affymetrix (contains the prefix, ALL) and cDNA (with the prefix, cDNA) expression levels, along with the Affy\(_{\text{ratio}}\) data for the seven different leukaemic children are plotted as shown in Figure 2. In comparison to the original Affymetrix (Affy\(_{\text{original}}\)) data, the plot indicates that the transformed Affymetrix, Affy\(_{\text{ratio}}\) aligns more closely with the cDNA than the Affy\(_{\text{original}}\).

Hierarchical clustering is useful to find the closest associations among gene profiles where it unsupervisedly seeks to build a hierarchy of clusters based on relatedness. Whether any unwanted change has been caused to the microarray data through the process of the transformation can be evaluated through hierarchical clustering. The method when applied to the pre- and post-transformed microarray data would highlight if any relative change has occurred to the state of the data. Accordingly, with Euclidean distance and Ward’s agglomerative
Figure 3. Hierarchical gene clustering of Affy_original (left) and Affy_ratio (right).

procedure\textsuperscript{23}, a divisive hierarchical clustering is conducted on the expression levels of Affymetrix genes before transformation and another on the expression levels. The result depicts that there is no significant variation between the two, as shown in Figure 3.

A similar hierarchical clustering is also applied to the patients to check whether the method has caused any change in the relationship among the patients. The outcome of this test too fails to substantiate that the modification caused to the data alters any relative relationship among the patients. Divisive hierarchical clustering of the patients is shown in Figure 4.

Both gene- and patient-clustering can be taken into consideration to support that the proposed approach does not upset the overall relationship in the microarray data. Next, it is required to evaluate the standing of this method in the midst of other microarray data merging approaches.

The process of sample standardization and gene centering is an approach, which is reportedly implemented as a data-merging process\textsuperscript{24–26}. This is used to compare the strength of our approach.

To implement sample standardization and gene centering, each microarray sample is first standardized; and, if there is variation in the range of data between the samples from both the platforms, then the gene was centred. However, it is difficult to judge how much variation would be considered appropriate; and therefore, gene centering is done once with sample standardization, and once without it.

In classical statistics, one of the fundamental distributions is the normal distribution or the Gaussian distribution. The probability density function for the normal distribution having mean, $\mu$ and standard deviation, $\sigma$ is given by eq. (2).

\begin{equation}
    f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}
\end{equation}

Each microarray sample from either platform can be standardized by making $\mu = 0$ and $\sigma = 1$ in the probability density function. This gives the probability density function for the standard normal distribution as shown in eq. (3).

\begin{equation}
    f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}
\end{equation}

Once the samples are standardized, each gene belonging to each study is centred. As the genes are arranged in the rows of the dataset while the columns contain the various samples, the gene centering is done by subtracting the row-wise mean from the values in each row of data so that the mean value of each row becomes zero. The samples from multiple platforms can be subsequently merged as sample standardization followed by centering of each gene in each study is done.

The method is applied to the normalized Affymetrix and cDNA data. The Pearson correlation coefficient is subsequently tested and found to be -0.02615. This explains that the sample standardization and gene centering method does not improve the correlation between the two datasets.

Further, gene centering alone is applied to the dataset of each platform. This time, however, the resultant correlation coefficient increases to 0.46. This implies that for the microarray data, the sample standardization is not required; instead only gene centering improves the relation. However, the value of this correlation still remains below the outcome obtained from the proposed method.
Distance weighted discrimination or DWD\textsuperscript{27} is a method used by Benito \textit{et al.}\textsuperscript{28} for batch correction and adjustments in biases including across microarray platform effects. It is based on modern statistical discrimination methods, and has reportedly been effective in removing biases present in a breast tumour microarray dataset. The method progresses by finding a direction, DWD direction, in which the sample-vectors from two studies are well-separated. It then translates the samples from each study along that direction until their respective families of vectors have significant overlap. This shifting of samples from each study in the DWD direction reportedly helps remove the biases. To evaluate the relative standing of our method, the DWD statistical correction algorithm is applied to the normalized datasets of Affymetrix and cDNA data belonging to the leukaemic patients. The test applying DWD method to our data produces a correlation of 0.77. Although the DWD method uses distance measures unlike our approach, there is an improvement in the latter method of merging the two sets of microarray data.

To compare further with other methods, approaches like XPN\textsuperscript{25} and probability of expression method\textsuperscript{29,30} have been explored. However, it is found that such methods are not suitable for relatively smaller sample size. Andrey Shabalin (pers. commun. to the authors) confirms in this regard that his team’s XPN method does not work for smaller sample size. This may again be argued as a negative aspect for such methods that they can only take large sample size into consideration. With regards to the gene centering and DWD methods, our approach can be ranked between these two.

As reviewed by Sarmah and Samarasinghe\textsuperscript{31}, there are various methods available for microarray data integration for large sample data. However, several of these methods, such as the probability of expression method\textsuperscript{29,30} and XPN\textsuperscript{25}, are found to be unusable for small microarray sample size. Besides being a non-complex exploit, the approach presented here can be applied to both small and large-sample data. Further, it works on the true expression measures unlike several other methods, where the core component in the data integration methodology involves transforming the data using measures, such as distance\textsuperscript{28,32}, probability scale\textsuperscript{29,30}, and ranking of fold change\textsuperscript{33,34}. While comparing the proposed approach with gene centering and the DWD methods, the latter provides a slight improvement over our method. However, there are a few virtues of the method proposed in this communication that deserve mention.

Our method is an attempt to view the problem of cross-platform data-integration from a distinctly different perspective, where the study focuses on the nature of the generated data from the individual platforms. The approach is crafted based on the fundamental characteristics of the two platforms and on the prominent distinguishing features of their relationship; and therefore, it has evolved from a sound base providing the required rigour. It also furnishes greater transparency as well as simpler applicability enabling a prospective user to relate to it depending on the basic knowledge about microarray technology in general while attaining similar or higher level of accuracy delivered by a variety of available complex statistical and machine-learning approaches. From this viewpoint, the proposed approach can counterbalance any apparent advantages of other available methods. In case of DWD, the method finds a separating hyperplane between the two microarray batches, adjusts the data by projecting the different batches on the DWD plane, finds the batch mean, and then subtracting out the DWD plane multiplied by this mean. With regards to this approach, Johnson and Li\textsuperscript{35} confirm that researchers face difficulties while trying to implement this method, and a few of the difficulties include that the method is ‘fairly complicated’ and can be applied to only two batches at a time. As an example of the DWD method, a stepwise approach was used by Benito \textit{et al.}\textsuperscript{28}, where they first adjusted the two most similar batches and then compared the third against the previous (adjusted) two. This stepwise method provided reasonable results in their three-batch case, but this could potentially break down in cases where there would be many more batches or when batches would not be very similar. Further, the DWD approach may also be considered as a black-box method, which tends to fall short of providing much insight into the process underneath.

The introduction and the subsequent rapid evolution of microarray technology has been a major step towards the
comprehensive biologic characterization of various cellular processes. The approach presented and discussed here highlights that its use can address the issue of incomparability of gene expression data from Affymetrix and cDNA microarray platforms. The work conducted maintains the highest housekeeping standards, and looks at the issue of microarray data integration from a different perspective. This approach is crafted based on the true gene expression array data integration from a different perspective. This platform highlights that its use can address the issue of incomparability.

RESEARCH COMMUNICATIONS


ACKNOWLEDGEMENT. The authors thank Dr Daniel Catchpoole, Head, Tumour Bank, The Children’s Hospital at Westmead, Australia for providing the data to conduct this study.

Received 8 March 2011; accepted 5 April 2012